ESIP Earth Science Data Analytics (ESDA) Cluster

August 21, 2014

Agenda

- **Items of Interest**
- **Frisco Recap**
- **Discussion: Descriptive Analytics**

Unfortunately, our get speaker had to postpone his talk to a future ESDA telecon

Relevant AGU Sessions

- Teaching Science Data Analytics Skills Needed to Facilitate Heterogeneous Data/Information Research: The Future Is Here - Session ID#: 1879
- Identifying and Better Understanding Data Science Activities, Experiences, Challenges, and Gaps Areas - Session ID#: 1809
- Advancing Analytics using Big Data Climate Information System - Session ID#: 3022
- Big Data in the Geosciences: New Analytics Methods and Parallel Algorithm - Session ID#: 3292
- Leveraging Enabling Technologies and Architectures to enable Data Intensive Science - Session ID#: 3041
- Open source solutions for analyzing big earth observation data - Session ID#: 3080
- Technology Trends for Big Science Data Management - Session ID#: 2525

2014 IEEE International Conference on Big Data (IEEE BigData 2014)

cci.drexel.edu/bigdata/bigdata2014/

Cal Tech – B...ta Analytics | NASA Science Plan 2014 | Main_Page | Data Science Central | eMOD– for GSFC visitors | toolmatch.esipfed.org | 2014 AGU Fall Meetin

What's New
Call for papers
Special Session
Sponsorship
Workshops
Online Submission
Highlights

In recent years, "Big Data" has become a new ubiquitous term. Big Data is transforming science, engineering, medicine, healthcare, finance, business, and ultimately society itself. The IEEE International Conference on Big Data 2014 (IEEE BigData 2014) provides a leading forum for disseminating the latest research in Big Data Research, Development, and Application.

We solicit high-quality original research papers (including significant work-in-progress) in any aspect of Big Data with emphasis on 5Vs (Volume, Velocity, Variety, Value and Veracity): big data science and foundations, big data infrastructure, big data management, big data searching and mining, big data privacy/security, and big data applications.

Frisco Recap - Agenda

- **Review:** What we have accomplished
- **Guest Speaker: Peter Fox** on the role of Data Scientist in facilitating the definition and subsequent usability of Data Analytics to enhance Earth science research
- **Summary of past speakers –**
  - Data Analytics needs and/or tools and their targets
  - Defining types of data analytics users http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics/Telecom_Presentations
- **Use Case Matrix Analysis** – Gleaning out Data Analytics needs http://wiki.esipfed.org/index.php/Use_Case_Collection
- **Data Analytics Tools Matrix** – What tools can provide appropriate analytics capabilities http://wiki.esipfed.org/index.php/Analytics_Tools

Frisco Recap - Observation

- **ESDA goal is to facilitate making information into knowledge**
- The ESDA Cluster, attracting a lot of interest, continues to 'churn' through the process of maturing their understanding and impacts of this new paradigm: Data Analytics and Data Science.
- Session participants were comprised of technologists and data users, with the majority of people, in attendance to 'learn'.
- Thus, in the early stages of this Cluster life, we continue to emphasize learning, which will doubtlessly evolve into applying (shaping) the knowledge we gain into implementable techniques that facilitate the use and advancement of data analytics and data science.

Frisco Recap – Guest Speaker Highlights

Peter Fox, Guest Speaker:

- Much of what is being done is relabeling and repacking- not much is being done new with new data science and analytics.
- GIS4Science and Data Analytics courses developed at RPI- there is no separate degree program in data science or informatics science – the courses are embedded in other programs (bioinformatics, physics, etc.)
- The Power in Analytics is Predictive and Prescriptive – in big data knowledge of nonparametric aspect is critical
- Students should be solving real problems with real science from the start, data science must be a skill – there is a key element of team work since data science is mainly done in groups.

Frisco Recap – Guest Speaker Highlights

Peter Fox, Guest Speaker:

- There are only 2 papers written on the theory of Data (one is from 1963) this makes it difficult to teach data and even more difficult for students to understand.
- It is important to distinguish between analytics and analysis – "Analysis is looking in (at the data) and Analytics is looking beyond (the data)"
- In discussing the scope of data analytics, from, for example, data discovery to science discovery, we need to be clear of the scope in which data analytics addresses. Thus we need to bound the problem being addressed.

Frisco Recap – Guest Speaker Highlights

- Science and interdisciplinary from the start!
- Not a question of: do we train scientists to be technical/ data people, or do we train technical people to learn the science
- It's a skill/ course level approach that is needed
- We teach methodology and principles over technology
- Data science must be a *skill*, and natural like using instruments, writing/using codes
- Team/ collaboration aspects are key
- Foundations and theory must be taught

Frisco Recap – Actions

- Flush out use cases: Bound the issue, be specific. Seek additional individuals who are facing issues utilizing large heterogeneous datasets
- Further define Data Analytics types: Per type: Issues, Potential solutions, exemplary situations, user classes, other
- Initiate some of the above planned mapping
- In January, have 2 ESDA sessions. One can be entitled: 'Earth Science Data Analytics 101'

Discussion - Descriptive Analytics

Descriptive Analytics: You can quickly understand "what happened" during a given period in the past and verify if a campaign was successful or not based on simple parameters.

Diagnostic Analytics: If you want to go deeper into the data you have collected from users in order to understand "Why some things happened," you can use … intelligence tools to get some insights.

Discoveritive Analytics: The use of data and analysis tools/models to discover information

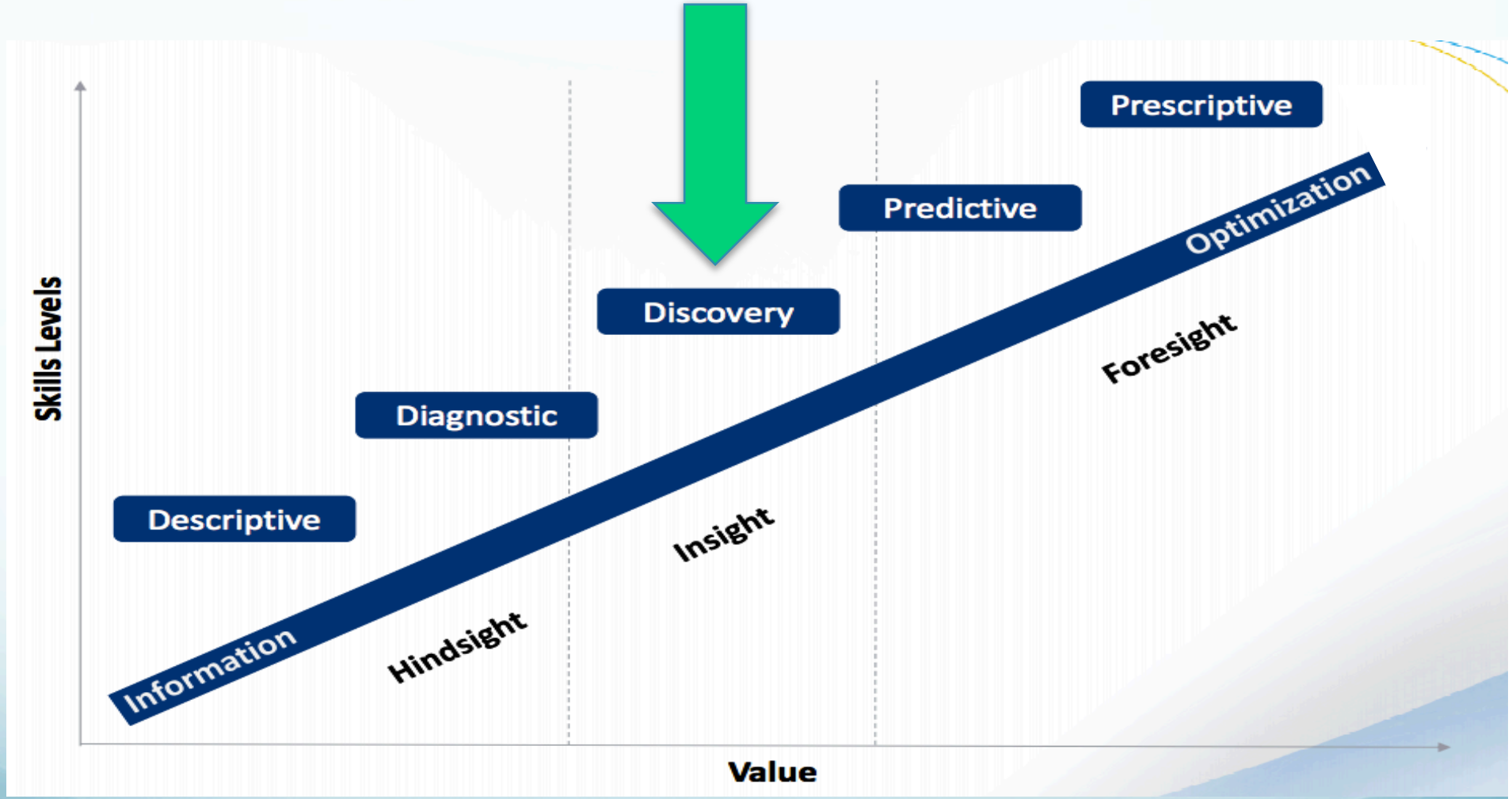
Predictive Analytics: If you can collect contextual data and correlate it with other user behavior datasets, as well as expand user data … you enter a whole new area where you can get real insights.

Prescriptive Analytics: Once you get to the point where you can consistently analyze your data to predict what's going to happen, you are very close to being able to understand what you should do in order to maximize good outcomes and also prevent potentially bad outcomes. This is on the edge of innovation today, but it's attainable!

Modified from: http://www.ciandt.com/card/four-types-of-analytics-and-cognition

Discovery Analytics:

This is where people learn from the data.



http://www.informationbuilders.es/intl/co.uk/presentations/four_types_of_analytics.pdf

Defining - Descriptive Analytics

Descriptive Analytics: You can quickly understand "what happened" during a given period in the past and verify if a campaign was successful or not based on simple parameters.

What does Descriptive Data Analytics mean? What does it do? How it is used? Examples! Where in Earth science would this be used? Which users?

- purpose of descriptive analytics is to summarize and tell you what has happened in the past
- "the simplest class of analytics," one that allows you to condense big data into smaller, more useful nuggets of information. *http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Reduction-2-Understanding-Predictive-Analytics/ba-p/79616*
- compute descriptive statistics (i.e. counts, sums, averages, percentages, simple arithmetic) that summarizes certain groupings or filtered version of the data, which are typically simple counts of some events. They are mostly based on standard aggregate functions *http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Reduction-1-Descriptive-Analytics/ba-p/77766*

Defining - Descriptive Analytics

Descriptive Analytics: You can quickly understand "what happened" during a given period in the past and verify if a campaign was successful or not based on simple parameters.

What does Descriptive Data Analytics mean? What does it do? How it is used? Examples! Where in Earth science would this be used? Which users?

- The purpose of descriptive analytics is simply to summarize and tell you what happened. For example, number of post, mentions, fans, followers, page views, kudos, +1s, check-ins, pins, etc. ...simple event counters.
- Other descriptive analytics may be results of simple arithmetic operations, such as share of voice, average response time, % index, average number of replies per post, etc.
  http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Reduction-1-Descriptive-Analytics/ba-p/77766
- Descriptive analytics is simple, all we need is data
- following the NetFlix approach, Amazon uses "Descriptive" analytics to process what you have purchased in the past, to predict what books, videos, and things you might like in the future

Defining - Descriptive Analytics

Descriptive Analytics: You can quickly understand "what happened" during a given period in the past and verify if a campaign was successful or not based on simple parameters.

What does Descriptive Data Analytics mean? What does it do? How it is used? Examples! Where in Earth science would this be used? Which users?

- Descriptive analytics answers the question, "What happened...?" It looks at data and information to describe the current situation in a way that trends, patterns and exceptions become apparent http://www.mu-sigma.com/analytics/ecosystem/dipp.html
- Descriptive statistics is the discipline of quantitatively describing the main features of a collection of information or the quantitative description http://en.wikipedia.org/wiki/Descriptive_statistics
- Natural Hazards: Looking for Patterns and Trends; Bringing in heterogeneous datasets, together summarized, to detect patterns
- Erin to provide slides: air quality 'use case'

User Model (Subsetted from ESDSWG WG)

Classes	Definition
Public	interested user of no or limited scientific skill
Graduate student	person of moderate to high skill at a university or college working towards an advanced degree
Production Centers	large organization that handles/processes vast quantities of data
Science Team	group of scientists focused on a specific area of study or on a specific instrument type, can include cal/val scientists
QA/Testing	developers or scientists using data to test software operation or to determine quality of a product, can include cal/val scientists
Data Analyst	person using NASA data to perform a specific analysis.
Domain Scientist	person using data to do research and publish within a discipline, comes in with some expertise in using the data
Interdisciplinary Scientist	person using high-level data products from multiple sources
Operational User	Data analyst or tech using data for operational support (applications) and emergency response
Assimilation Modelers	persons or groups that routinely obtain vast quantities of data for incorporation into models, can have operational needs